Biometrics 60, 116–123
March 2004

Estimation of Fish Abundance Indices Based on Scientific Research Trawl Surveys

Jiahua Chen,* Mary E. Thompson, and Changbao Wu

Department of Statistics and Actuarial Science, University of Waterloo,
Waterloo, Ontario N2L 3G1, Canada
**email:* jhchen@uwaterloo.ca

Summary. The fish abundance index over an ocean region is defined here to be the integral of expected catch per unit effort (CPUE), approximated by the sum of expected CPUE over grid squares. When trawl surveys are done within grid squares selected according to a probability sampling design, several other sources of variation such as the fish population dynamics and the catching process are also involved. In such situations model-assisted methods for estimating abundance, assessed under both design and model perspectives, have some advantages over purely design-based methods such as the Horvitz–Thompson (HT) estimator or purely model-based prediction approaches. This article develops model-assisted empirical likelihood (EL) methods via loglinear regression and nonparametric smoothing. The methods are applied to grid surveys of the Grand Bank region carried out annually by Fishery Products International from 1996 through 2002. The HT and EL methods produce similar point estimates of abundance indices. Simulation results, however, indicate that the EL estimator under local linear smoothing is associated with smaller standard errors.

Key words: Auxiliary information; Empirical likelihood method; Loglinear model; Model-assisted approach; Nonparametric smoothing; Survey design; Variance estimation.

1. Introduction

One problem of key importance to the management of fishery resources is to obtain a reliable estimate of fish abundance. The absolute fish abundance in an open ocean region is hard to obtain. Instead, we often settle on estimating abundance indices. Current practice in managing marine fisheries in the area of the Northwest Atlantic Fisheries Organization (NAFO) involves setting annual quotas for commercial species with high abundance and maximum allowable percentages of by-catch for species with low abundance which are under monitoring. The estimated fish abundance indices, along with others, are the major factors used in the decision-making process.

Before the 1970s, fish abundance indices were estimated using the catch-effort data reported by commercial fishing units. It gradually became evident that such estimates are not reliable, as many factors, including vessel and gear type, crew experience, and underreporting, are confounded with the abundance estimates. In addition, commercial fishing activities inevitably focus on high population density areas and therefore paint a biased abundance picture. Scientific research trawl surveys using a standard vessel and gear type and a probability sampling design have been adopted by many organizations since the 1970s (Doubleday and Rivard, 1981). They allow us to define a more objective abundance index and to obtain more accurate estimates for the abundance indices and for the population distributions of the entire region as well.

We define the response variable Y in a fishery trawl survey as the number (or the biomass) of fish caught in a given location by a research vessel with standard fishing gear through a unit fishing time, i.e., catch per unit effort (CPUE). Suppose a random set of locations is selected and the corresponding response variables are observed. There are three main sources of variation in the data. The first is the dynamics of the fish population: The number of fish at the given location changes constantly over time; the second is the fishing process: Only a random subset of the total fish at the location is captured; the third is the randomness of the location: Only a random subset of feasible locations determined by the survey design is observed.

In fishery literature, data from scientific research trawl surveys are often analyzed using conventional model-based or design-based methods in survey sampling. The model-based approach makes inference according to an assumed probability model for the response variable, ignoring the randomness associated with the selection of sampling units; the design-based approach makes inference according to the randomness induced by the sampling design, treating the response variable as nonrandom for each of the units.

In this article, we build a framework where two sources of randomness, the catching process and the probability selection of sampling units, are entertained. Our method is similar to the model-assisted approach in survey sampling but the resulting estimator is assessed using the model and the sampling design jointly. This is different from the

conventional model-assisted approach, which in general uses a design-based analysis. We consider a short period of time such that the fish population can be viewed as fixed. We postulate a semiparametric or nonparametric model that relates the expected CPUE to various hydrographic and bathymetric variables and fit the model based on data available from the survey. We define a conceptually related abundance index and then estimate the index using a model-assisted approach. In Section 2, we formally define the fish abundance index under the current framework. Issues and methods related to model building based on survey data are discussed in Section 3. Model-assisted estimators are introduced in Section 4. Variance estimators for the estimated abundance indices are presented in Section 5. In Section 6, we apply our proposed methodology to grid survey data of the Grand Bank region (NAFO divisions 3LNO) provided by Fishery Products International (FPI) Ltd. of Canada. Also in Section 6, a simulation study has been conducted to investigate the small sample performances of three estimators considered in the application, using a synthetic population created from the FPI survey data. We conclude with some brief remarks in Section 7.

2. The Fish Abundance Index

A fish population in a large open area does not constitute a conventional finite population. It is a mobile population that changes over time due to migration, recruitment, natural mortality, and fishing mortality. The fundamental principle for design-based analysis in survey sampling, namely that the finite population parameters are fixed and can be determined without error by conducting a census, is grossly violated. The underlying population dynamics are hard to observe and the true stock size may never be known. Reed (1986) describes this as a "black box" situation. Fortunately, what is really important for stock assessment and management is to monitor the fluctuation of the fish population so that a major decline or boost in the population size or the total population biomass can be detected, and consequently appropriate management strategies can be adopted.

The response variable CPUE should not be used directly to form the abundance index as it is likely affected by many known or unknown factors. In addition to the fish population size in the given region and time period, the CPUE may also depend on, among other things, the fishing vessel and gear type, the towing speed, the location, the time of the day, the temperature of the water, and the roughness of the ocean conditions. Despite all of these caveats, however, it may still be reasonable to believe that there exists a conceptually expected CPUE value as a smooth function of location and the aforementioned covariates in a given short period of time of the year. We aim to define the abundance index as the integral of expected CPUE value over the region at standardized levels of important covariates.

Let R denote the region where abundance of a certain fish is of interest. Let $\mu(\boldsymbol{x})$ be the expected CPUE at $\boldsymbol{x} \in R$ under standard conditions at the given time period. Mathematically, our definition is as follows.

DEFINITION 1: *The fish abundance index in the region R is defined as* $I(R) = \int_R \mu(\boldsymbol{x})\, d\boldsymbol{x}$.

This definition, however, is hard to use in practice. In addition to the difficulty in specifying a functional form for $\mu(\boldsymbol{x})$, it is computationally awkward when covariates other than locations are also involved. One way to circumvent these shortcomings is to form an index as follows. We divide the region into N equal-sized grid squares represented by g_i, $i = 1, 2, \ldots, N$. These grid squares should be large enough to accommodate, say, a 30-minute tow by a typical fishing vessel. Within each grid square g_i, a response variable Y_i could be thought of as a CPUE obtained by "standard" fishing gear under "standard" ocean conditions. Note that Y_i is random, and we assume that there is an unspecified probability model ξ behind this randomness. Under this model, we assume $E_\xi(Y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i)$ where $\boldsymbol{x}_i$ consists of covariates (including location) associated with grid square g_i. This discretization is equivalent to assuming that the expected CPUE value is constant within each grid square. With this convention, our practical definition of fish abundance index is as follows.

DEFINITION 2: *The fish abundance index in the region R is computed as* $I(R) = \sum_{i=1}^{N} \mu(\boldsymbol{x}_i)$.

This index is not fish population size itself but a conceptual and discretized expectation, which provides a relative measurement of the true fish abundance. More importantly, it is possible to construct unbiased estimators for the abundance index under this framework while it is almost impossible to do so for the true population abundance. In fishery literature, it is often postulated that $\mu(\boldsymbol{x}_i) = q\lambda(\boldsymbol{x}_i)$, where q is the so-called catchability coefficient (see, for example, Schnute, 1994) and $\lambda(\boldsymbol{x}_i)$ is the true stock size in the ith grid square. If q can be determined from other sources, then $\hat{I}(R)/\hat{q}$ serves as an estimate for the total stock size. The true relation between $\mu(\boldsymbol{x}_i)$ and $\lambda(\boldsymbol{x}_i)$, however, can be very complicated (Gunderson, 1993). In this article, we focus on estimating the index $I(R)$ using scientific research trawl survey data. The issue of how such an index is related to the actual population abundance will not be addressed.

In grid surveys, it is Y_i, not $\mu(\boldsymbol{x}_i)$, that is observed at each sampled grid square g_i. There are two sources of randomness here, the random selection of sampling units (grid squares) and the random observation of catch over a unit effort. An estimator of $I(R)$ should be assessed with respect to both the model, ξ, and the design, p. To avoid confusion, we clarify our interpretation of unbiasedness with the following definition.

DEFINITION 3: *Let $\hat{I}(R)$ be an estimator of $I(R)$. We term $\hat{I}(R)$ an unbiased estimator of $I(R)$ if* $E\{\hat{I}(R)\} = E_\xi E_p\{\hat{I}(R)\} = I(R)$.

3. Models for the Catching Process

Fitted values for $\mu(\boldsymbol{x}_i)$ are required for estimating the abundance index $I(R)$ and can be obtained through a working model.

3.1 *The Loglinear Model*

An overdispersed loglinear model can be a natural choice for catch data. The model is semiparametric and is specified through the first and the second order moments of Y_i given the $\boldsymbol{x}_i$, i.e., $\log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, $V_\xi(Y_i \mid \boldsymbol{x}_i) = \sigma^2\mu_i$, where $\mu_i = \mu(\boldsymbol{x}_i) = E_\xi(Y_i \mid \boldsymbol{x}_i)$ and V_ξ denotes the variance under

the catch model. The σ^2 is often referred to as overdispersion parameter. For catch data it is most likely that $\sigma^2 > 1$.

Estimation of the model parameters $\boldsymbol{\beta}$ and σ^2 based on the survey data may follow the general framework of Godambe and Thompson (1986). Let s be the set of labels for the sampled grid squares, n be the sample size, and $\pi_i = P(i \in s)$ be the inclusion probabilities from the survey design. Let $d_i = 1/\pi_i$ be the design weights. The quasi-maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is the solution to $U(\boldsymbol{\beta}) = \boldsymbol{X}'_n \boldsymbol{W}_n(\boldsymbol{Y}_n - \boldsymbol{\mu}_n) = \mathbf{0}$, where $\boldsymbol{x}_n = (\boldsymbol{x}'_1, \ldots, \boldsymbol{x}'_n)'$, $\mathbf{Y}_n = (y_i, \ldots, y_n)'$, $\boldsymbol{\mu}_n = (\mu_1, \ldots, \mu_n)'$, $\mu_i = \exp(\boldsymbol{x}'_i\boldsymbol{\beta})$, and $\mathbf{W}_n = \mathrm{diag}(d_1, \ldots, d_n)$. The y_i's are the observed values for the Y_i's. The fitted values for $\mu_i = \mu(\boldsymbol{x}_i)$ are given by $\hat{\mu}_i = \exp(\boldsymbol{x}'_i\hat{\boldsymbol{\beta}})$. A moment estimator for σ^2 can then be obtained based on Pearson-type fitted residuals: $\hat{\sigma}^2 = (N-k)^{-1}\sum_{i\in s} d_i(y_i - \hat{\mu}_i)^2/\hat{\mu}_i$, where k is the dimension of the parameter space for $\boldsymbol{\beta}$.

3.2 *Nonparametric Smoothing*

Under the assumption that $\mu(\boldsymbol{x})$ is a smooth function of $\boldsymbol{x}$, we could estimate $\mu(\boldsymbol{x}_i)$ by a weighted mean of sampled y_j over the whole region. The general form is $\hat{\mu}(\boldsymbol{x}_i) = \sum_{j\in s} w(\boldsymbol{x}_j, \boldsymbol{x}_i)y_j$. There exists a large variety of weighting schemes under various smoothing strategies in the literature. Our experience with fishery applications shows that the local linear estimator is an ideal one to use in terms of the overall performance of the abundance index estimators. Another advantage of local linear smoothing is the reduced boundary bias (Fan and Gijbels, 1996), which could be substantial for other methods under two or higher dimensional kernel analysis.

The local linear estimator estimates $\mu(\boldsymbol{x}_i)$ by fitting a linear model in a neighborhood of $\boldsymbol{x}_i$. However, the $\hat{\mu}(\boldsymbol{x}_i)$ can also be expressed as a linear function of the y_j's. When $\boldsymbol{x}_i = (x_{1i}, x_{2i})'$ is of dimension two, it can be shown that $w(\boldsymbol{x}_j, \boldsymbol{x})$ is proportional to

$$
\begin{aligned}
&\sum_{i\in s} x_{1i}^2 K_h(\boldsymbol{x}_i - \boldsymbol{x}) \sum_{i\in s} x_{2i}^2 K_h(\boldsymbol{x}_i - \boldsymbol{x}) \\
&\quad - \left\{ \sum_{i\in s} x_{1i}x_{2i} K_h(\boldsymbol{x}_i - \boldsymbol{x}) \right\}^2 + \{x_{1j} K_h(\boldsymbol{x}_j - \boldsymbol{x})\} \\
&\quad \times \left\{ \sum_{i\in s} x_{1i}x_{2i} K_h(\boldsymbol{x}_i - \boldsymbol{x}) \sum_{i\in s} x_{2i} K_h(\boldsymbol{x}_i - \boldsymbol{x}) \right. \\
&\qquad \left. - \sum_{i\in s} x_{1i} K_h(\boldsymbol{x}_i - \boldsymbol{x}) \sum_{i\in s} x_{2i}^2 K_h(\boldsymbol{x}_i - \boldsymbol{x}) \right\} \\
&\quad + \{x_{2j} K_h(\boldsymbol{x}_j - \boldsymbol{x})\} \\
&\quad \times \left\{ \sum_{i\in s} x_{1i} K_h(\boldsymbol{x}_i - \boldsymbol{x}) \sum_{i\in s} x_{1i}x_{2i} K_h(\boldsymbol{x}_i - \boldsymbol{x}) \right. \\
&\qquad \left. - \sum_{i\in s} x_{1i}^2 K_h(\boldsymbol{x}_i - \boldsymbol{x}) \sum_{i\in s} x_{2i} K_h(\boldsymbol{x}_i - \boldsymbol{x}) \right\}.
\end{aligned}
$$

It is then rescaled so that $\sum_{j\in s} w(\boldsymbol{x}_j, \boldsymbol{x}) = 1$.

Two important issues in nonparametric smoothing are the choices of the kernel function K and the bandwidth parameter h. It is widely accepted that the choice of K is not crucial. The literature on the choice of h is extensive. Many existing methods, such as plug-in, cross-validation, and generalized cross-validation, have equal convergence rates but differ in terms of asymptotic variances (Härdle, Hall, and Marron, 1988; Ruppert, Sheather, and Wand, 1995; Wand, and Jones, 1995). In this article, we use the data-driven cross-validation method which minimizes the total prediction error $\sum_{i\in s}\{\hat{\mu}_{-i}(\boldsymbol{x}_i) - y_i\}^2$, where $\hat{\mu}_{-i}(\boldsymbol{x}_i)$ is computed with $(y_i, \boldsymbol{x}_i)$ being removed from the data set. This method is simple to implement and has good properties in terms of mean integrated squared errors (Xia and Li, 2002).

4. Methods of Estimation

In this section, we first provide a brief review of conventional model-based and design-based estimation methods and then present our proposed model-assisted estimators for fish abundance indices. We show that the maximum pseudo-empirical likelihood estimator is particularly appealing in the current context. The method is not only highly efficient but very flexible for incorporating auxiliary information as well as historical survey data through the assumed catching model. It is also robust against model misspecifications. In what follows we consider a well-defined region R, and the abundance index is simply denoted by I.

4.1 *The Design-Based and the Model-Based Approaches: A Review*

There are two conventional approaches in the fishery literature for analyzing catch-effort data. See, for example, Smith (1990) for a discussion. In the current context, the sample data are in the form of $\{(y_i, \boldsymbol{x}_i), i \in s\}$. The "design-based" Horvitz–Thompson (HT) estimator

$$\hat{I}_{HT} = \sum_{i\in s} d_i y_i$$

for I is widely used, where $d_i = 1/\pi_i$ are the design weights. Note that we use p to denote the randomization distribution induced by the probability sampling design. Under the design-based framework,

$$E_p(\hat{I}_{HT}) = \sum_{i=1}^{N} y_i,$$

where for $i \notin s$ the y_i is conceived as the actual catch should the ith grid square be sampled. It is clear that only if one treats the observed y_i as the true $\mu(\boldsymbol{x}_i)$ can this estimator be viewed as design-unbiased for the abundance index I of Definition 2. This amounts to ignoring the variation associated with the catching process. The estimator is unbiased according to Definition 3, but it ignores all the auxiliary information that is routinely collected from the survey.

The model-based approach assumes that the amount of catch Y from a unit effort is random and follows a parametric probability distribution. The abundance index I is defined and estimated through the estimation of model parameters. The fact that a portion of the observed catches are zero, while the rest take positive values and are right-skewed, leads to the proposal of using the so-called Δ distribution in estimating the abundance index. This distribution assigns a positive parameter for the probability of having zero catch, and often a lognormal distribution for the nonzero catches (Pennington,

1983; Smith, 1988). The mean parameter of the Δ distribution is treated as the abundance index and is estimated through the maximum likelihood method. This approach can also be extended to incorporate auxiliary information. Another version of the model-based approach is to rewrite $\sum_{i=1}^{N} y_i$ as $\sum_{i\in s} y_i + \sum_{i\notin s} y_i$ and to use a prediction type of estimator $\sum_{i\in s} y_i + \sum_{i\notin s} \hat{y}_i$ through an assumed model (Smith, 1990).

The use of a parametric model such as the Δ distribution can provide efficient estimates for the abundance indices if the model is appropriate. More often than not, however, the assumed model is too restrictive to describe the particular situation under study and results in severe bias in estimating the target abundance indices. In addition, the nature of complex survey data makes model building and diagnostics difficult tasks to fulfill. Due to these considerations, the purely model-based approach will not be included in the FPI application presented in Section 6.

4.2 *The Model-Assisted Estimators*

There have been three main model-assisted estimators proposed in the survey research literature. These estimators can be applied under the current context to obtain more efficient and robust estimators for the abundance indices. The one with the simplest structure is the generalized difference (GD) estimator (Cassel et al., 1976).

$$\hat{I}_{\mathrm{GD}} = \sum_{i\in s} d_i y_i - \sum_{i\in s} d_i \hat{\mu}_i + \sum_{i=1}^{N} \hat{\mu}_i.$$

The GD estimator is more efficient than the HT estimator (i.e., has smaller mean squared error) when $\hat{\mu}(\boldsymbol{x}_i)$ is a good enough predictor of Y_i that the residual variable $r_i = y_i - \hat{\mu}(\boldsymbol{x}_i)$ has a smaller design-based variance than that of the variable y_i itself. This estimator is robust against model misspecifications in that $E(\hat{I}_{\mathrm{GD}}) = E_\xi E_p(\hat{I}_{\mathrm{GD}}) \doteq I$ regardless of the working model used to obtain $\hat{\mu}(\boldsymbol{x}_i)$.

The model-calibration method (Wu and Sitter, 2001) can also be applied here to obtain a generalized regression (GR) estimator for the abundance index by treating $\hat{\mu}_i$ as an auxiliary variable:

$$\hat{I}_{\mathrm{GR}} = \sum_{i\in s} d_i y_i + \hat{B}\left(\sum_{i=1}^{N} \hat{\mu}_i - \sum_{i\in s} d_i \hat{\mu}_i\right),$$

where $\hat{B}$ is the estimated regression coefficient of y_i on μ_i. Since letting $\hat{B} = 1$ reduces the GR estimator to the GD estimator, an optimal choice of $\hat{B}$ guarantees some improvement of GR over GD in theory. The GR estimator is also approximately unbiased for I irrespective of the working model used for the catching process.

The third and more recently proposed pseudo-empirical likelihood (EL) method (Chen and Sitter, 1999; Wu and Sitter, 2001) can ideally be used for the estimation of I. Note that the GR estimator can be rewritten as a weighted sum in the form $\sum_{i\in s} w_i y_i$. The EL estimator also shares this form, i.e., $\hat{I}_{\mathrm{EL}} = \sum_{i\in s} w_i y_i$, with $w_i = Np_i$ and p_i being the maximizer of the pseudo-empirical loglikelihood function

$$l(\boldsymbol{p}) = \sum_{i\in s} d_i \log(p_i)$$

subject to constraints

$$\sum_{i\in s} p_i = 1 (p_i > 0) \quad \text{and} \quad \sum_{i\in s} p_i \hat{\mu}_i = \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}_i.$$

The EL estimator compares favorably with the GR estimator in many aspects and has several distinctive features not enjoyed by the other two methods.

For many commonly used sampling designs, such as the stratified random sampling commonly used in fish abundance surveys, and under some mild regularity conditions on the modeling, the EL estimator is asymptotically equivalent to the GR estimator (Wu and Sitter, 2001). One can make either choice, $\hat{I}_{\mathrm{GR}}$ or $\hat{I}_{\mathrm{EL}}$, without any loss of efficiency or robustness. The GR estimator, when written as $\hat{I}_{\mathrm{GR}} = \sum_{i\in s} w_i y_i$, has an undesirable property: the weight w_i can be negative. This drawback is inherent to a regression-type estimator. Theoretically, it is possible to have a negative estimate $\hat{I}_{\mathrm{GR}}$ for the index I. The weights $w_i = Np_i$ for the EL estimator, on the other hand, are always positive and hence guarantee nonnegative estimation for I. The EL estimator has a clear maximum likelihood interpretation, a feature that is often preferred by practitioners. In terms of computation, simple and stable algorithms for computing the EL weights are available (Chen et al., 2002).

The biggest advantage of using the EL estimator, however, is the natural extension of the method to incorporating historical data or data from other sources to improve the abundance estimates. It has been observed from fishery history that a fish population usually evolves slowly over time. A sudden dramatic change in the total stock size or major migration of the entire fish school from one region to another is unlikely. Historical data or data from other sources collected from the same region can be very valuable.

Let $\mu_i(t)$ be the expected grid CPUE for year t, with $t = 2$ representing the current year and $t = 1$ the past year (or other time period). Let $\hat{\mu}_i(2)$ be obtained using the current year data and $\hat{\mu}_i(1)$ using historical data or data from other sources. Let $\{y_i,\, i \in s\}$ be the observed CPUE from the current year survey. The EL estimator of I which uses historical data as additional auxiliary information is computed as $\hat{I}_{\mathrm{EL}} = N\sum_{i\in s} p_i y_i$, where the p_i maximize $l(\boldsymbol{p})$ subject to

$$\sum_{i\in s} p_i = 1 (p_i > 0) \quad \text{and} \quad \sum_{i\in s} p_i \hat{\mu}_i(t) = \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}_i(t),\ t = 1, 2.$$

The effect of adding one more constraint into the maximization process is equivalent to including an extra independent variable in regression analysis. The estimator using both constraints ($t = 1, 2$) will perform asymptotically at least as well as the one using a single constraint ($t = 2$). The gain in efficiency depends only on the correlation of data between this year and past years. No models are needed to relate the data explicitly over different years. The basic properties of the EL estimator are preserved even if the historical data are irrelevant. In such situations the resulting estimator might be less stable when the sample size is small, a scenario similar to that of the generalized regression estimator. In addition, one can incorporate data from many years or many sources by simply adding more constraints. It is not a problem if the data from

other sources are collected using different vessels or fishing gear.

There are many existing methods that employ models built on biological constraints, or models that accommodate historical data to some degree. Most of those models are state space models (Schnute, 1994) or autoregressive models (Roff, 1983) that explicitly relate the data structure over the time sequence. Efficient analysis from those models can be expected when the assumed model is correct, but substantial bias could also be induced when the model is misspecified.

5. Variance Estimation

The two sources of variation associated with the catching process and the sampling selection induce two variance components for the estimated abundance indices. Let Var denote the overall variance under the mixed ξp randomization, and V_ξ and V_p the variances under the model and the survey design, respectively. For the "design-based" HT estimator, we have

$$\begin{aligned}\mathrm{Var}(\hat{I}_{\mathrm{HT}}) &= V_\xi\{E_p(\hat{I}_{\mathrm{HT}})\} + \mathrm{E}_\xi\{V_p(\hat{I}_{\mathrm{HT}})\} \\ &= V_\xi\left(\sum_{i=1}^{N} y_i\right) + E_\xi\{V_p(\hat{I}_{\mathrm{HT}})\}.\end{aligned}$$

The first term is the variance component due to the randomness in fishing, and the second is due to probability sampling. Typically, the first term is of order $O(N)$ and the second is of order $O(N^2/n)$. In cases where the sampling fraction is negligible, i.e., $n/N \doteq 0$, estimating $\mathrm{Var}(\hat{I}_{\mathrm{HT}})$ is asymptotically equivalent to estimating $V_p(\hat{I}_{\mathrm{HT}})$.

For the grid survey defined in this article, since a single grid square needs to be defined large enough to accommodate a typical tow at a certain speed over a 30- or 60-minute time period, the total number of grid squares (N) over a fixed region is usually moderate. For instance, the stratum sampling fraction used in the FPI survey is sometimes as large as 80%. Variance estimators following the so-called "design-based" approach by ignoring the first component will undoubtedly underestimate the true variance, resulting in unreliable confidence intervals for the abundance indices.

The first component cannot be estimated without using a model. Under the loglinear model discussed in Section 3, assuming that y_i, $i = 1, 2, \ldots, N$ are conditionally independent given $\boldsymbol{x}_i$, $i = 1, 2, \ldots, N$, we have

$$V_\xi\left(\sum_{i=1}^{N} y_i\right) = \sigma^2 \sum_{i=1}^{N} \mu_i = \sigma^2 I.$$

An approximately unbiased variance estimator for $\hat{I}_{\mathrm{HT}}$ is given by

$$\mathrm{Var}(\hat{I}_{\mathrm{HT}}) = \hat{\sigma}^2 \hat{I} + v_p(\hat{I}_{\mathrm{HT}}),$$

where $v_p(\hat{I}_{\mathrm{HT}})$ is a design unbiased estimator of $V_p(\hat{I}_{\mathrm{HT}})$.

The empirical likelihood estimator $\hat{I}_{\mathrm{EL}}$ is a nonlinear estimator and its exact variance does not have a closed form. Under certain regularity conditions on the sampling design and on the assumed model, similar to those used by Chen and Sitter (1999) and Wu and Sitter (2001), we can show that

$$\hat{I}_{\mathrm{EL}} = \hat{I}_{\mathrm{HT}} + B\left(\sum_{i=1}^{N} \mu_i - \sum_{i\in s} d_i \mu_i\right) + o_p(Nn^{-1/2}),$$

where $B = (\sum_{i=1}^{N} y_i\mu_i)/(\sum_{i=1}^{N} \mu_i^2)$, and the stochastic order o_p is with respect to the probability sampling. It follows immediately that $E_p(\hat{I}_{\mathrm{EL}}) = E_p(\hat{I}_{\mathrm{HT}}) + o(Nn^{-1/2})$, and consequently $\mathrm{Var}(\hat{I}_{\mathrm{EL}}) \doteq V_\xi(\sum_{i=1}^{N} y_i) + E_\xi\{V_p(\hat{I}_{\mathrm{EL}})\}$, where the design-based variance component is given by

$$V_p(\hat{I}_{\mathrm{EL}}) \doteq V_p\left\{\sum_{i\in s} d_i(y_i - B\mu_i)\right\}.$$

For most commonly used sampling designs, we have $V_p(\hat{I}_{\mathrm{EL}}) < V_p(\hat{I}_{\mathrm{HT}})$ which implies $\mathrm{Var}(\hat{I}_{\mathrm{EL}}) < \mathrm{Var}(\hat{I}_{\mathrm{HT}})$. Estimation of this design-based variance component and construction of a variance estimator for the EL estimator $\hat{I}_{\mathrm{EL}}$ are straightforward. Confidence intervals for the abundance indices can be constructed by resorting to the usual normal approximations.

6. Application to the FPI Grid Survey of the Grand Bank Region

Grid surveys have been conducted over part of the Grand Bank region on the east coast of Canada since 1996 by FPI Ltd. The survey aims to provide estimates of abundance indices for several important commercial fish species including yellowtail flounder, Atlantic cod, and American plaice. Our analysis reported here is based on the 2001 survey and uses the data for yellowtail flounder. More analyses on the FPI grid survey data can be found in a detailed technical report available from the authors.

The entire region is divided into 626 equal-sized grid squares, each 10 × 10 square miles. A stratified random sampling design was used where strata boundaries were determined based on practicality in terms of data collection as well as homogeneity of fish abundance within each stratum. The grid map and the strata boundaries are shown in Figure 1. An approximately optimal sample size allocation scheme was used where stratum sample size is proportional to the estimated total abundance index of the stratum computed using data from other sources.

6.1 *Modeling*

We first build a model that includes all important auxiliary variables. Preliminary examination of the data reveals that variables that should be considered include x_1 and x_2: the latitude and longitude of the location; x_3: light during the tow; x_4: time of the day; and x_5: average depth of the tow. The response variable Y is the number of fish caught per minute during a 30-minute tow.

Due to economic and physical constraints over trawl surveys, the tows were arranged so that nearby grid squares are more likely to be towed in an ordered fashion. We may hence expect some temporal or sequential effect in the survey data. However, a preliminary time series analysis shows that the temporal effect is very minor once the spatial factors are incorporated into the model. That is, the temporal and spatial effects are confounded and including the spatial covariates into the model will likely remove most of the temporal effect.

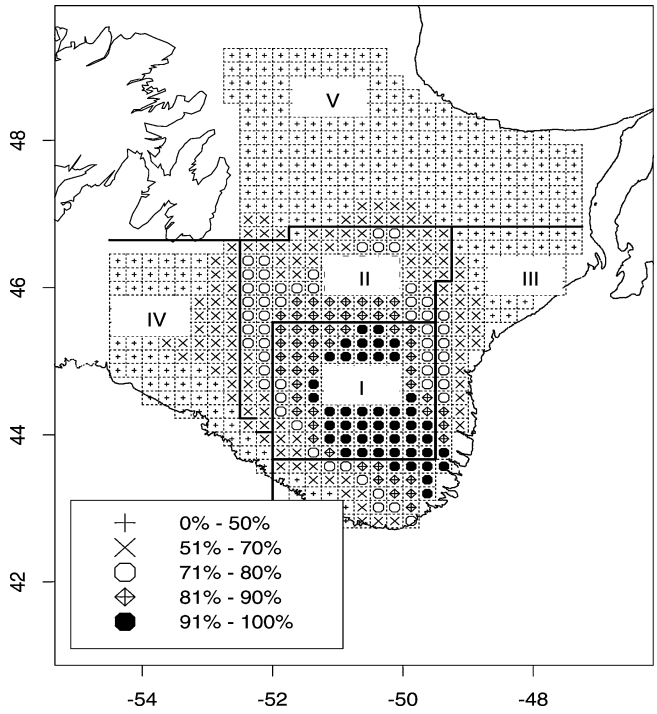

Figure 1. Grid map and population density plot.

For the purpose of predicting the expected CPUE over individual grid squares, the conditional model assuming the Y_i's are independent given the $\boldsymbol{x}_i$'s seems plausible. Note that this conditional independence assumption is used only for variance estimation.

For the loglinear model, it is determined that terms of second order and/or interaction between x_1 and x_2 may also be important. Results from statistical hypothesis tests show that both x_1 and x_2 and their second order terms as well as the interaction term are significant; x_3 is also significant but x_4 and x_5 are not, provided that x_1, x_2, and x_3 are included in the model. Our final loglinear model includes x_1, x_2, x_1^2, x_2^2, $x_1 \times x_2$, and x_3.

For the local linear regression method, only x_1 and x_2 are included in the model. We suspect that not imposing a restrictive form for $\mu(\boldsymbol{x})$ allowed x_1 and x_2 to have better interpretive power than they have in the loglinear model. This judgment seems to be supported by the simulation results reported in Section 6.3.

Our estimation procedure requires $\boldsymbol{x}$ values at each grid square in the entire region. We do have the first two components x_1 and x_2 corresponding to the central location of each grid square. Information on other $\boldsymbol{x}$ components were collected only for the sampled grid squares. In the loglinear model we set the third covariate x_3 to its median value 4.5 (x_3 was recorded using a 0–9 scale) in computing the predicted value $\hat{\mu}(\boldsymbol{x}_i)$. With this adjustment the impact of x_3 was appropriately taken into account.

6.2 *The Estimated Abundance Indices over the Grand Bank Region*

We computed three abundance index estimates using the HT estimator ($\hat{I}_{\mathrm{HT}}$), the EL estimator under the loglinear model ($\hat{I}_{\mathrm{EL1}}$), and the EL estimator under local linear smoothing ($\hat{I}_{\mathrm{EL2}}$). The three estimators provide similar point estimates for the abundance index, with $\hat{I}_{\mathrm{EL2}}$ having the smallest estimated standard error for almost all cases.

Table 1
Estimated abundance indices for the Grand Bank region

Stratum	$\hat{I}$	%	(SE)
1	2917.0	49.4	(176.4)
2	1688.3	28.7	(143.4)
3	997.8	16.9	(127.1)
4	230.2	3.9	(68.8)
5	63.9	1.1	(28.8)
Total	5889.8	100	(271.1)

Our results reported in Table 1 are based on $\hat{I}_{\mathrm{EL2}}$, and the overdispersion parameter σ^2 is estimated using the combined data from all strata. A very large proportion (78.1%) of the total abundance is distributed over the first two strata (I and II), with some fish activity in stratum III. The large northern area (stratum V) has virtually no presence of this species.

We supplement Table 1 with the smoothed population density plot as shown in Figure 1. The abundance index at a given grid square is the predicted value $\hat{\mu}(\boldsymbol{x}_i)$ based on local linear smoothing. The percentages shown are the population abundance index quantiles for the entire region. It is clear that stratum I contains almost exclusively the top 10% population quantiles, and strata I and II include most of the top 30% population quantiles, with some minor indication that certain fish schools appear close to the deep water of the southeast part of the region.

6.3 *A Simulation*

While theoretical considerations and the results from Section 6.2 are in favor of the EL estimator under local linear smoothing, it is difficult to draw conclusions regarding the finite sample performance of the estimator based only on a single analysis without knowing the true abundance index. The extremely high sampling fractions in strata I and II also undermine the performance of the EL estimator since in such cases the model variance component is likely to prevail. These issues, however, can be investigated through simulation studies.

To reduce the huge computational burden over repeated simulation runs where loglinear model fitting, cross-validation for local linear smoothing, and maximum empirical likelihood estimation have to be carried out for each of the simulated samples, we consider a smaller region consisting of $N = 200$ grid squares from the original strata I and II. However, stratification will not be used in the simulation. The very high sampling fractions over this region also enable us to create a "true index" that is close to the real world. The values of $\mu(\boldsymbol{x}_i)$ are obtained through local linear smoothing, and these smoothed $\mu(\boldsymbol{x}_i)$ are treated as the true expected CPUE; consequently the true index $I = \sum_{i=1}^{N} \mu(\boldsymbol{x}_i)$ is known under this setting. Only the location variables are included as covariates.

At each simulation run, the response variable Y is first generated by using a negative binomial distribution such that

Table 2
Simulated variance components for abundance index estimators ($\times 10^3$)

		$\sigma^2 = 1.5$			$\sigma^2 = 6.0$		
f	$\hat{I}$	V_1	V_2	MSE	V_1	V_2	MSE
0.3	*HT*	7.36	71.92	79.28	30.24	125.12	155.36
	*EL*1	7.36	38.04	45.40	30.24	97.72	127.96
	*EL*2	7.36	29.24	36.60	30.24	96.56	126.80
0.5	*HT*	7.36	29.12	36.48	31.00	53.16	84.16
	*EL*1	7.36	14.76	22.12	31.00	39.92	70.92
	*EL*2	7.36	10.16	17.52	31.00	36.68	67.68
0.7	*HT*	7.32	12.00	19.32	29.60	18.96	48.56
	*EL*1	7.32	6.92	14.24	29.60	12.36	41.96
	*EL*2	7.32	4.36	11.68	29.60	12.28	41.88

$E_\xi(Y_i \,|\, \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i)$ and $V_\xi(Y_i \,|\, \boldsymbol{x}_i) = \sigma^2\mu(\boldsymbol{x}_i)$ for a prechosen overdispersion parameter, σ^2. A sample s of n grid squares is then drawn by simple random sampling without replacement. The three estimators mentioned in Section 6.2 are computed for each sample. This process is repeated $B = 1000$ times. The simulated bias and mean squared error for an estimator $\hat{I}$ are computed as $\text{Bias}(\hat{I}) = B^{-1}\sum_{b=1}^{B}(\hat{I}_b - I)$ and $\text{MSE}(\hat{I}) = B^{-1}\sum_{b=1}^{B}(\hat{I}_b - I)^2$, respectively, where $\hat{I}_b$ denotes the estimate $\hat{I}$ from the bth simulated sample. The model variance component is given by $V_1 = B^{-1}\sum_{b=1}^{B}(\sum_{i=1}^{N} Y_{bi} - I)^2$ where Y_{bi} is the response variable from the ith grid square and the bth simulation run. The variance component due to sampling is obtained as $V_2 = \text{MSE} - V_1$ assuming the bias is negligible.

Results for three choices of sample size and two values of the overdispersion parameter are presented in Table 2. The relative biases $|\text{Bias}(\hat{I})/I|$ for all cases are less than 0.5% and thus are not reported. The three sample sizes represent typical sampling fractions at low ($f = 0.3$), median ($f = 0.5$), and high ($f = 0.7$) levels in trawl surveys, and the two values of σ^2 correspond to mild overdispersion ($\sigma^2 = 1.5$) and severe overdispersion ($\sigma^2 = 6$).

The simulation results can be summarized as follows: (1) the EL estimator under local linear smoothing ($\hat{I}_{\text{EL2}}$) has the smallest overall variance (or MSE) among all three estimators considered; (2) the EL estimator under a loglinear model ($\hat{I}_{\text{EL1}}$) has performance close to $\hat{I}_{\text{EL2}}$ in many cases but never outperforms $\hat{I}_{\text{EL2}}$; (3) the HT estimator ($\hat{I}_{\text{HT}}$) has substantially larger overall variance (or MSE) when the sampling fraction is not too large; and (4) the model variance component becomes dominant when the sampling fraction is high, particularly under severely overdispersed models.

7. Concluding Remarks

Under our model-assisted framework used in this article, fish abundance surveys are not viewed as classical design-based finite population problems. There are two major sources of variation that should be considered at both the sampling design stage and the estimation stage, namely, the variation due to the catching process and the variation due to the sampling design. The associated variance components are very different in nature and both play important roles in the estimation of fish population abundance indices.

Standardization is the key to reducing the variance component due to fishing. The FPI grid survey, which uses a standard scientific vessel, *Atlantic Lindsey*, and a standard fishing gear with a group of experienced crew members at a fixed time period of the year, is an excellent move in that direction. Information on factors that could affect the fishing outcome such as temperature and sea bottom type can be used to build more accurate models for the fishing process. This information is not collected by the current FPI survey. Such information would be even more useful if collected for all grid squares in the region, not just sampled grid squares.

An optimal survey design and suitable estimation techniques are crucial to reducing the other variance component due to sampling. A near-optimal stratified design should be attractive for practical considerations when strata boundaries are chosen for the convenience of the actual trips for data collection. The empirical likelihood method is very attractive in terms of estimation. Auxiliary information and historical data can easily be used to improve the abundance estimates.

Acknowledgements

This research was supported by a subvention research grant from the Department of Fisheries and Oceans of Canada. We wish to thank the Fishery Products International Ltd. of Canada for providing us the grid survey data. Thanks are also due to the associate editor and a referee for constructive comments and suggestions, which led to substantial improvement of the article.

Résumé

L'indice d'abondance de poisson pour une région océanique est défini ici comme l'intégrale de la prise moyenne par tentative (CPUE moyenne), qu'on approxime par la somme des CPUE espérées sur un maillage. Lorsque les recensements au chalut sont effectués dans des maillages sélectionnés selon un schéma d'échantillonnage probabilisé, plusieurs autres sources de variation interviennent, telles que les dynamiques des populations de poisson, et le processus de pêche. Dans ces situations, des méthodes assistées par modèle pour estimer l'abondance, reposant à la fois sur le schéma et sur le modèle, sont avantageuses par rapport à des méthodes basées seulement sur le schéma comme l'estimateur de Hurvitz-Thomson (HT) ou seulement sur une approche de prédiction par modèle. Ce papier développe des méthodes, assistées par modèle, de vraisemblance empirique (EL) par régression non linéaire et lissage non paramétrique. Les méthodes sont appliquées aux recensements maillés de la zone de pêche du Grand Banc réalisés chaque année de 1996 à 2002 par Fishery Products International. Les méthodes HT et EL donnent des estimations ponctuelles des indices d'abondance similaires. Des résultats de simulation indiquent cependant que l'estimateur EL avec lissage local linéaire est associé à de plus petits écarts-type.

References

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and

generalized regression estimation for finite populations. *Biometrika* **63,** 615–620.

Chen, J. and Sitter, R. R. (1999). A pseudo-empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica* **9,** 385–406.

Chen, J., Sitter, R. R., and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika* **89,** 230–237.

Doubleday, W. G. and Rivard, D. (1981). *Bottom Trawl Surveys*. Canadian Special Publication of Fisheries and Aquatic Sciences 58.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. New York: Chapman & Hall.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review* **54,** 127–138.

Gunderson, D. R. (1993). *Surveys of Fisheries Resources*. New York: Wiley.

Härdle, W., Hall, P., and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *Journal of the American Statistical Association* **83,** 86–95.

Pennington, M. (1983). Efficient estimators of abundance for fish and plankton surveys. *Biometrics* **39,** 281–286.

Reed, W. J. (1986). Analyzing catch-effort data allowing for randomness in the catching process. *Canadian Journal of Fisheries and Aquatic Science* **43,** 174–186.

Roff, D. A. (1983). Analysis of catch/effort data: A comparison of three methods. *Canadian Journal of Fisheries and Aquatic Science* **40,** 1496–1506.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90,** 1257–1270.

Schnute, J. T. (1994). A general framework for developing sequential fisheries models. *Canadian Journal of Fisheries and Aquatic Science* **51,** 1676–1688.

Smith, S. J. (1988). Evaluating the efficiency of the Δ-distribution mean estimator. *Biometrics* **44,** 485–493.

Smith, S. J. (1990). Use of statistical models for the estimation of abundance from groundfish trawl survey data. *Canadian Journal of Fisheries and Aquatic Science* **47,** 894–903.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. New York: Chapman & Hall.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association* **96,** 185–193.

Xia, Y. and Li, W. K. (2002). Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *Journal of Multivariate Analysis* **83,** 265–287.

Received January 2003. *Revised September* 2003.
Accepted September 2003.